https://eventos.utfpr.edu.br//sicite/sicite2019

Validação de dados do Repositório Brasileiro Livre para Dados Abertos do Solo

Data validation of the Free Brazilian Repository for Open Soil Data

Mikael Pereira Messias
mikaelmessias@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Santa Helena, Paraná, Brasil

Alessandro Samuel-Rosa
alessandrorosa@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Santa Helena, Paraná, Brasil

RESUMO

O objetivo deste trabalho é apresentar uma solução automatizada para encontrar inconsistências nos dados do Repositório Brasileiro Livre para Dados Abertos do Solo (febr). O febr possui um manual que define todos os padrões do repositório, como convenções de codificação e unidades de medida, e é imprescindível que os conjuntos de dados estejam nos padrões lá definidos para garantir o bom funcionamento do repositório. A estrutura do repositório consiste em planilhas eletrônicas armazenados no Google Drive, serviço de armazenamento em nuvem oferecida pela Google. A Google oferece também o Apps Script, uma plataforma de desenvolvimento cuja linguagem é baseada na versão 3 do ECMAScript que permite criar novas funcionalidades para os aplicativos do Google. Foi desenvolvido um complemento para o Google Sheets, que verifica automaticamente o tipo de tabela do conjunto e procura por inconsistências nos dados, adicionando notas onde o responsável pelo conjunto deve realizar correções. Embora o complemento tenha sido apenas parcialmente implementado, as validações das tabelas 'dataset' e 'observacao' já estão disponíveis para uso final.

PALAVRAS-CHAVE: Repositórios institucionais. Google Apps. Ciência do solo.

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

Direito autoral: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.

ABSTRACT

The aim of this paper is to present an automated solution to find inconsistencies in data from the Free Brazilian Repository for Open Soil Data (febr). Since febr has a manual that defines all repository standards, such as coding conventions and units of measurement, it is indispensable that all datasets are in the standards defined there to ensure the proper functioning of the repository. The repository structure consists of spreadsheets stored in Google Drive, Google's cloud storage service. Google also offers Apps Script, a development platform whose language is based on ECMAScript version 3 that lets you create new functionality for Google applications. A Google Sheets add-on has been developed that automatically checks the table type and looks for data inconsistencies by adding notes where the dataset manager should make corrections. Although the add-on has been partially implemented, both 'dataset' and 'observation' tables validation are now available for end use.

KEYWORDS: Institutional repositories. Google Apps. Soil science.

INTRODUÇÃO

No meio científico, é comum que novos experimentos utilizem dados obtidos por outros pesquisadores para produzir novos resultados. Portanto, é importante que os dados produzidos estejam estruturados e organizados. Por outro lado, ao considerar a utilização de dados obtidos por outras pessoas também é necessário considerar a padronização e harmonização desses dados. Esses aspectos afetam diretamente a capacidade de reusabilidade dos dados.

A ciência do solo é uma das áreas que sofrem pela falta de organização adequada dos dados gerados e de meios padronizados para acessá-los, seja por falta de incentivo, de domínio de ferramentas tecnológicas ou de padrões comuns. Mesmo sendo uma área que já produziu e ainda produz muitos dados, existe certa dificuldade para obtê-los, muitas vezes cabendo aos interessados realizar a compilação de todos os dados que precisa, um processo que pode ser moroso dependendo das barreiras que surgirem. Esse tipo de esforço pode ser minimizado através da utilização de bases de dados centralizadas que armazene todos os dados de forma organizada e pronta para o uso. Nesse sentido, existem esforços ao redor do mundo que buscam criar bases de dados científicos para os mais diversos fins. Naturalmente cada base de dados apresenta padrões diferentes para os dados recebidos, mas a ideia principal é a mesma: receber dados de várias fontes, organizar esses dados e encaixá-los em padrões que sejam aceitos pela comunidade científica.

A título de exemplo, o *World Soil Information Service* (WoSIS), criado e mantido pelo *International Soil Reference and Information Centre* (ISRIC), é uma base de dados que oferece dados do solo padronizados e harmonizados do mundo inteiro. Os dados submetidos para o WoSIS são importados em um banco de dados PostgreSQL do jeito que foram recebidos, assim uma cópia dos dados originais fica armazenada no repositório. Só então os dados são importados ao WoSIS, onde são mapeados para o padrão de dados do repositório e harmonizados (quando possível) e distribuídos, podendo ser utilizados para gerar produtos a serem apresentados para a comunidade de usuários (RIBEIRO; BATJES; OOSTRUM, 2018). O WoSIS possui um rigoroso processo de padronização, realizando o maior esforço possível para que todos os dados estejam mapeados para um único padrão, incluindo as convenções de nome e os valores das propriedades do solo. Um outro exemplo de base de dados é o Pangaea, que arquiva, publica e distribui dados georreferenciados de pesquisas do sistema terrestre. O Pangaea, ao contrário do WoSIS, apresenta pouca ou nenhuma padronização nos dados submetidos ao repositório.

O Repositório Brasileiro Livre para Dados Abertos do Solo (febr) surgiu em 2018 como uma parceria entre a Universidade Federal de Santa Maria (UFSM) e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), um esforço coletivo de cientistas do solo com o objetivo de armazenar e compartilhar todo e qualquer tipo de dado do solo do Brasil. O febr aposta no uso de métodos que buscam o equilíbrio entre as restrições do WoSIS e a flexibilidade do Pangaea. Os dados são armazenados em formato de planilhas, validados com relação aos padrões do repositório, harmonizados e por fim incluídos em um superconjunto de dados

que pode ser utilizado para os mais diversos fins. A validação e harmonização desses dados atualmente é realizada por estudantes e professores que fazem parte da equipe febr. Esse processo pode levar dias, semanas ou até meses, dependendo de fatores como a quantidade de dados presentes nos conjuntos e da disponibilidade dos materiais de onde esses dados foram retirados.

O objetivo deste trabalho é apresentar uma solução automatizada para encontrar inconsistências e validar os dados do conjuntos de dados do febr, principal produto desenvolvido durante a vigência do projeto.

Material e Métodos

O desenvolvimento do projeto foi divido em duas etapas principais. A primeira etapa baseou-se primeiramente na familiarização com a ciência do solo e dados do solo, obtida com aprovação na disciplina de gênese e morfologia do solo ofertada pelo curso de Agronomia da Universidade Tecnológica Federal do Paraná (UTFPR), câmpus Santa Helena. Ao mesmo tempo era necessária a familiarização com a estrutura do repositório, e para alcançar isso foram executadas verificações manuais de inconsistências nos dados em vários conjuntos de dados do febr. Essas verificações eram realizadas com o objetivo de encontrar dados que não harmonizavam com os padrões do repositório, como unidades de medida, códigos inexistentes ou não reconhecidos, campos em branco, entre outros.

A segunda etapa foi baseada no desenvolvimento dos *scripts* de validação dos dados. As regras de padronização de dados do manual do febr foram resgatadas e divididas entre dois conjuntos: básica e extra. A validação básica inclui as regras que garantem o funcionamento básico do febr, enquanto a validação extra inclui regras mais abrangentes, mas que não são necessárias para o funcionamento mais básico do febr. A Figura 1 apresenta um exemplo de regras de validação da tabela 'dataset'.

Figura 1 – Regras de validação

Regras de validação
Tabela dataset

Item	Descrição	Tipo
Todos os campos	Verificar se os códigos dos campos são exatamente idênticos aos códigos presentes na planilha de padrões. A exceção é o campo `dataset_referencia_i`, que pode aparecer múltiplas vezes, onde o índice `i` deve ser igual a um número inteiro indicando a ordem de importância da referência, não sendo permitida sua repetição.	Básica
Todos os campos	Verificar se todos os campos possuem algum valor especificado, ou seja, nenhuma célula em branco.	Extra
Campo `dataset_id`	Verificar se o valor é composto pelo prefixo `ctb` seguido de um número inteiro de quatro dígitos, por exemplo, `0001`.	Básica

Fonte: Autoria própria (2019).

Os padrões do febr são baseados em experiências internacionais e registrados num manual de compilação, organização e revisão de dados

(SAMUEL-ROSA et al, 2019). O manual descreve os padrões de codificação, unidades de medida, estrutura das tabelas, entre outros detalhes inerentes ao bom funcionamento do repositório. Quando essas convenções não são utilizadas, podem surgir diversas inconsistências nos dados que precisam ser corrigidas para manter o bom funcionamento do febr.

Como os conjuntos de dados são armazenados no formato de planilhas eletrônicas em diretórios do Google Drive, existem ferramentas que podem ser utilizadas em conjunto com os aplicativos Google. Uma delas é o Google Apps Script, uma plataforma de desenvolvimento lançada em 2009 e incorporada aos aplicativos Google que permite estender e adicionar novas funcionalidades à suíte de aplicativos (FERREIRA, 2014). A linguagem de desenvolvimento é um dialeto desenvolvido pelo Google baseado na versão 3 do ECMAScript, mas os *scripts* são executados nos servidores da empresa ao invés do navegador como acontece com o JavaScript, que também segue especificações do ECMAScript (MCPHERSON, 2016).

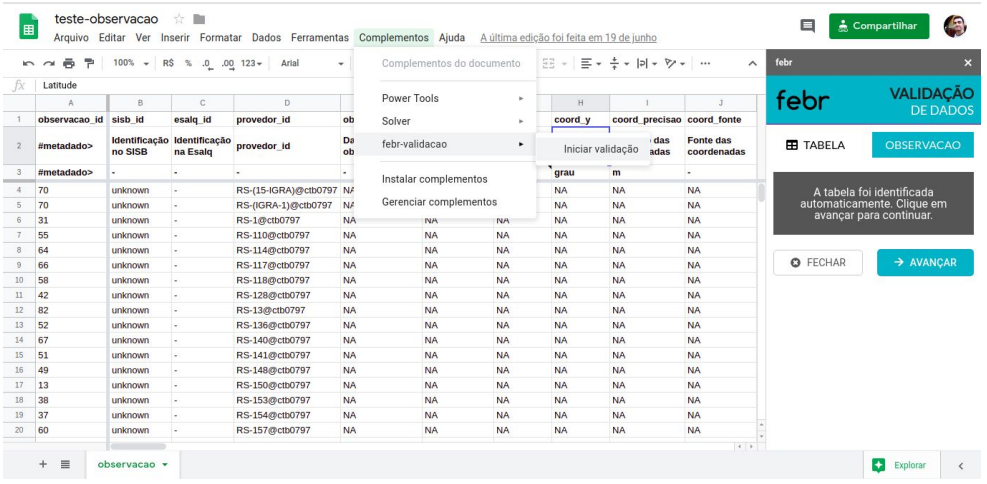
O Google Apps Script permite criar soluções utilizando os recursos fornecidos pelos aplicativos Google, com a possibilidade de criar até mesmo aplicações inteiramente novas usando os recursos providos pela Google. A execução das aplicações pode ser realizada diretamente da plataforma de desenvolvimento ou através de interfaces de usuário. Todos os arquivos gerenciados pelos aplicativos Google, com exceção dos arquivos do Google Slides e do Google Drawings, podem ser manipulados através do Apps Script, que transforma estes arquivos em objetos manipuláveis através de serviços como o *DriveApp* e *SpreadsheetApp* (MCPHERSON, 2016). Existem três formas de criar uma interface de usuário com essa ferramenta. A principal delas é chamada *container-bound*, e é exibida dentro do documento como uma janela *pop-up* ou uma barra lateral. A interface pode ser desenvolvida utilizando o serviço HTML do Apps Script, o que permite que sejam utilizadas tecnologias como jQuery e CSS, oferecendo maior liberdade e flexibilidade (FERREIRA, 2014).

Resultados e Discussão

As verificações realizadas de forma manual resultaram em algumas perspectivas interessantes sobre a organização e a disponibilidade dos relatórios originais e seus respectivos dados. Embora alguns relatórios estivessem disponíveis na internet, outros nem mesmo estavam digitalizados. Nesses casos, era necessário alcançar os autores e a instituição onde o relatório havia sido publicado para conseguir acesso ao mesmo. Além disso, em alguns conjuntos surgiram dados coletados e descritos de forma subjetiva, que precisaram ser conferidos com os autores do relatório quando possível, ou discutidos com outros cientistas do solo. Esse tipo de situação se demonstrou frequente principalmente em trabalhos mais antigos, onde era comum que não fossem esclarecidos os critérios e métodos utilizados, ou descritos de maneira incompleta (como a falta de coordenadas espaciais). Nesse momento, começaram a surgir questionamentos quanto aos procedimentos que deveriam ser tomados, uma vez que era necessário decidir se esses conjuntos seriam ignorados ou se seriam forçados a encaixar-se nos padrões do repositório.

O complemento foi implementado utilizando a solução *built-in* do Google Apps Script, surgindo como uma opção disponível na própria planilha e exibindo uma janela lateral quando ativado, como demonstrado na Figura 2.

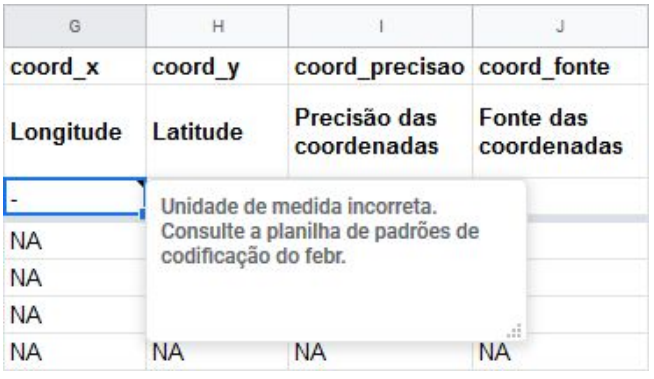
Figura 2 – Complemento do febr aberto como uma janela do Google Sheets



Fonte: Autoria própria (2019).

O complemento verifica automaticamente a tabela de acordo com o nome dado a ela, bastando o usuário seguir os passos para iniciar a validação. Ao iniciar a execução, os campos delimitados nas regras serão verificados quanto aos padrões do febr. As notas são recursos usados para identificar as células onde existem inconsistências exibindo mensagens customizadas ao usuário. A Figura 3 demonstra uma nota gerada pelo complemento.

Figura 3 – Inconsistência encontrada em uma das células



Fonte: Autoria própria (2019).

Conclusão

A princípio, percebe-se que é necessário manter os dados antigos, geralmente são mais problemáticos, da forma que foram disponibilizados inicialmente. Por outro lado, é necessário que os padrões e a estrutura do febr sejam evoluídos constantemente para atender estes dados, uma vez que o ponto

forte do repositório e o diferencial com relação às demais opções de repositórios disponíveis é sua flexibilidade. O desafio é lidar com a dependência entre o manual e o complemento, uma vez que o complemento precisará de atualizações sempre que surgirem novos padrões no manual do febr.

A porção implementada do complemento atingiu o resultado esperado. Ainda é necessário implementar soluções para a validação das tabelas 'camada' e 'metadado'. Entretanto, a validação das tabelas 'dataset' e 'observacao' através do complemento foram realizadas com sucesso, e já se apresenta como uma primeira versão viável. A expectativa é que o complemento seja finalizado e publicado com mais um ano de trabalho. Além disso, é necessário que sejam realizados testes com usuários finais (cientistas do solo) que possam validar a qualidade e utilidade do complemento. Com a publicação do complemento, espera-se passar gradativamente para os produtores dos dados a responsabilidade de manter os dados consistentes, até que não seja mais necessária interferência dos mantenedores do febr nos conjuntos.

Outro ponto a ser analisado após a finalização do complemento é relacionado ao potencial que este tem para atingir novos rumos através da execução de tarefas mais específicas, como a validação de coordenadas espaciais através de conexões com outros serviços, como o Google Maps. Embora o desafio seja grande, é algo que pode beneficiar muito o repositório e os cientistas do solo em busca de dados íntegros e de qualidade.

AGRADECIMENTOS

A esta universidade e a Fundação Araucária, pelo apoio e pelo fomento da pesquisa.

REFERÊNCIAS

RIBEIRO, E.; BATJES, N. H.; OOSTRUM, A. v. **World Soil Information Service (WoSIS)** - Towards the standardization and harmonization of world soil data. Procedures manual 2018, Report 2018/01, ISRIC – World Soil Information. Wageningen, 2018. Disponível em: <https://www.isric.org/sites/default/files/isric_report_2018_01.pdf>. Acesso em: 08 jul. 2019.

SAMUEL-ROSA, A; TEIXEIRA, W. G.; VIANA, J. H. M.; ROCHA, A. S.; GUBIANI, P. I; GRIS, D. J.; ROSIN, N. A. **Repositório Brasileiro Livre para Dados Abertos do Solo**: Manual de compilação, organização e revisão de dados. Disponível em: <https://docs.google.com/document/d/1Bqo8HtitZv11TXzTviVq2bI5dE6_t_fJt0HE-l3IMqM>. Acesso em: 29 ago. 2019.

FERREIRA, J. **Google Apps Script**: Web Application Development Essentials. 2. ed. Sebastopol: O'Reilly, 2014.

MCPHERSON, B. **Going GAS**: From VBA to Google Apps Script. Sebastopol: O'Reilly, 2016.